Joint imputation procedures for categorical variables with application to the French Wealth Survey

Hélène CHAPUT (*), Guillaume CHAUVET(**), David HAZIZA(***), Laurianne SALEMBIER (*), Julie SOLARD (*),

(*) Insee
(**) Crest, Ensai
(***) Université de Montréal

1 Introduction

Single imputation, which consists of replacing a missing value by an artificial value, is often used in statistical agencies for treating item nonresponse. The main objective of imputation is to reduce the nonresponse bias, which may be appreciable when the respondents and the nonrespondents differ with respect to the study variables. Key to reducing the nonresponse bias is the availability of powerful auxiliary variables for all the sample units (respondents and nonrespondents). Household and social surveys typically collect categorical variables. In order to avoid the possibility of impossible values in the imputed data file, it is customary to use some form of donor imputation methods such as nearest-neighbour imputation or random hot-deck imputation. This type of imputation consists of selecting (at random or not) a respondent (donor) from the set of respondents and using the donor's item values to "fill in" the missing value of a nonrespondent (recipient). In this paper, we focus on survey weighted random hot-deck imputation, under which donors are selected at random with probability proportional to the sampling weight. In practice, survey weighted random hot-deck imputation is generally applied within imputation classes, which are formed on the basis of auxiliary information recorded for the sample units (respondents and nonrespondents).

Most often, survey statisticians are interested in estimating simple parameters such as population totals or means. In this case, marginal imputation, which consists of imputing variables separately, leads to asymptotically unbiased estimators, provided the assumed model is correctly specified (Haziza, 2009). For example, one may use random hot-deck imputation for each variable requiring imputation. However, marginal imputation distorts the relationship between variables. As a result, estimators of bivariate parameters such as regression and correlation coefficients may be severely biased, especially if the nonresponse rates are appreciable. Thus, it is desirable to develop imputation strategies which succeed in preserving the relationship between categorical variables. For bivariate parameters involving continuous variables, Shao and Wang (2002) proposed a joint random regression imputation procedure and showed that it leads to asymptotically unbiased estimators of correlation coefficients. Chauvet and Haziza (2011) proposed a fully efficient version of the Shao-Wang procedure in the sense that the imputation variance is eliminated or at least, considerably reduced. A different approach for dealing with bivariate parameters was considered in Skinner and Rao (2002), who proposed to first use marginal imputation to fill in the missing values and then to adjust for the bias at the estimation stage.

In this paper, we propose a simple joint random imputation procedure that is closely related to random hot-deck imputation. We show that the proposed procedure preserves the correlation coefficient between two categorical variables. For simplicity, we consider the case of binary variables but the extension to the case of more than two categories is relatively straightforward.

We illustrate the proposed methods in the context of the French Wealth Survey (FWS), which is conducted by the French National Institute of Statistics and Economic Studies (Insee) every six years since 1986. The FWS collects information on many aspects of wealth : financial assets, real-estate assets, business wealth, but also social, cultural and symbolic capital. Also, information concerning the households (e.g., number of individuals in the household, age, occupation, income) is collected in order to understand the origins of wealth. In this paper, we focus on the 2010 FWS, for which the fieldwork began on October 2009 and ended in March 2010. Approximately 20,000 households, belonging to the metropolitan area, the French West Indies and the Reunion Island, were selected to be part of the survey.

2 Set-up

Consider a finite population U of size N. Let x and y denote two study variables such that $x_i = 1$ if unit i possesses characteristic A and $x_i = 0$, otherwise, and y is similarly defined with A replaced by some other characteristic B. We are interested in estimating the finite population correlation coefficients between x and y:

$$\rho_{xy} = \frac{t_{11} - t_{10}t_{01}/N}{\left(t_{10} - t_{10}^2/N\right)^{1/2}\left(t_{01} - t_{01}^2/N\right)^{1/2}},$$

where $t_{kl} = \sum_{i\in U} x_i^k y_i^l$, $(k,l) \in \{(1,0),(1,1),(0,1)\}$. Note that t_{10} (respectively, $p_{10} = t_{10}/N$) represents the number (respectively, the proportion) of individuals in the population who possess the characteristic A. Similarly t_{01} (respectively, $p_{01} = t_{01}/N$) represents the number (respectively, the proportion) of individuals in the population who possess the characteristic B. Finally, t_{11} (respectively, $p_{11} = t_{11}/N$) represents the number (respectively, the proportion) of individuals in the population who possess both characteristics.

A sample s is selected from U according to a sampling design $p(s)$. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)'$ be the vector of sample selection indicators, where $\delta_i = 1$ if unit i is selected in the sample and $\delta_i = 0$, otherwise. Let $w_i = 1/\pi_i$ be the sampling weight attached to unit i, where $\pi_i = P(i \in s)$ denotes its first-order inclusion probability in the sample. A complete data estimator of ρ_{xy} is the plug-in estimator given by

$$\hat{\rho}_{xy\pi} = \frac{\hat{t}_{11,\pi} - \hat{t}_{10,\pi}\hat{t}_{01,\pi}/\hat{N}_\pi}{\left(\hat{t}_{10,\pi} - \hat{t}_{10,\pi}^2/\hat{N}_\pi\right)^{1/2}\left(\hat{t}_{01,\pi} - \hat{t}_{01,\pi}^2/\hat{N}_\pi\right)^{1/2}}, \tag{1}$$

where $\hat{t}_{kl,\pi} = \sum_{i\in s} w_i x_i^k y_i^l$ is the expansion estimator of t_{kl}, and $\hat{N}_\pi = \sum_{i\in s} w_i$ is the expansion estimator of the population size N. The complete data estimator (1) is asymptotically unbiased for ρ_{xy}; e.g., Deville (1999). We denote by $\hat{p}_{10,\pi} = \hat{t}_{10,\pi}/\hat{N}_\pi$ the estimated proportion of individuals who possess the characteristic A. The estimated proportions $\hat{p}_{01,\pi} = \hat{t}_{01,\pi}/\hat{N}_\pi$ and $\hat{p}_{11,\pi} = \hat{t}_{11,\pi}/\hat{N}_\pi$ are similarly defined.

In practice, both x and y are prone to missing values and some form of imputation is required. We adopt the following notation: let r_{xi} be a response indicator attached to unit i such that $r_{xi} = 1$ if i responds to item x and $r_{xi} = 0$, otherwise. Similarly, let $r_{yi} = 1$ if i responds to item y and $r_{yi} = 0$, otherwise. We denote by $\mathbf{r} = (r_{x1}, ..., r_{xN}, r_{y1}, ..., r_{yN})'$ the vector of response indicators. Let x_i^* be the imputed value used to replace the missing x_i and y_i^* be the imputed value corresponding to missing y_i. Finally, let $\tilde{x}_i = r_{xi}\, x_i + (1-r_{xi})x_i^*$ and $\tilde{y}_i = r_{yi}\, y_i + (1 - r_{yi})y_i^*$. An imputed estimator of ρ_{xy} is given by

$$\hat{\rho}_{xyI} = \frac{\hat{t}_{11,I} - \hat{t}_{10,I}\hat{t}_{01,I}/\hat{N}_\pi}{\left(\hat{t}_{10,I} - \hat{t}_{10,I}^2/\hat{N}_\pi\right)^{1/2}\left(\hat{t}_{01,I} - \hat{t}_{01,I}^2/\hat{N}_\pi\right)^{1/2}}, \tag{2}$$

where $\hat{t}_{kl,I} = \sum_{i\in s} w_i \tilde{x}_i^k \tilde{y}_i^l$. We denote by $\hat{p}_{10,I} = \hat{t}_{10,I}/\hat{N}_\pi$ the imputed estimator of the proportion of individuals who possess the characteristic A. The imputed estimators $\hat{p}_{01,I} = \hat{t}_{01,I}/\hat{N}_\pi$ and $\hat{p}_{11,I} = \hat{t}_{11,I}/\hat{N}_\pi$ are similarly defined. Note that, once the data have been imputed, the computation of (2) does not require the response flags to be available in the imputed data file. In other words, complete data estimation procedures may be readily applied by secondary analysts. This is an important practical aspect since, in many situations, the response flags are not available in the files. If the response flags are available, an alternative to (2) is the complete case estimator, which is based on the units that responded to both items. The latter estimator is included in the empirical study presented in Section 4.

In this paper, we study the properties of $\hat{\rho}_{xyI}$, under the so-called nonresponse model (NM) approach. Let $P(r_{xi} = 1, r_{yi} = 1) \equiv p_{rr}$, $P(r_{xi} = 1, r_{yi} = 0) \equiv p_{rm}$, $P(r_{xi} = 0, r_{yi} = 1) \equiv p_{mr}$ and $P(r_{xi} = 0, r_{yi} = 0) \equiv p_{mm}$. Also, we assume that the sample units respond independently of one another. Let $\mathbf{x} = (x_1, ..., x_N)'$ and $\mathbf{y} = (y_1, ..., y_N)'$, where x_i and y_i denote the i-th value corresponding to items x and y, respectively. Under the NM approach, we define the conditional nonresponse bias of $\hat{\rho}_{xyI}$ as

$$B_{qI}(\hat{\rho}_{xyI}) = E_q E_I \left\{(\hat{\rho}_{xyI} - \hat{\rho}_{xy\pi}|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}\right\},$$

where the subscripts q and I denote respectively the unknown nonresponse mechanism and the imputation mechanism used for the random selection of donors. To simplify the notation, we write $E_I\left(\hat{\rho}_{xyI}|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}\right) \equiv \tilde{\rho}_{xyI}$ in the remainder of the paper.

In order to study the properties of $\hat{\rho}_{xyI}$, we express its total error as:

$$\hat{\rho}_{xyI} - \rho_{xy} = (\hat{\rho}_{xy\pi} - \rho_{xy}) + (\tilde{\rho}_{xyI} - \hat{\rho}_{xy\pi}) + (\hat{\rho}_{xyI} - \tilde{\rho}_{xyI}). \tag{3}$$

The first term on the right hand side of (3) represents the sampling error, whereas the second and the third terms represent the nonresponse error and the imputation error, respectively. Note that the imputation error occurs solely from the random selection of donors (see Section 3).

To obtain an asymptotically unbiased estimator of ρ_{xy}, we seek an imputation procedure under which $B_{qI}(\hat{t}_{kl,I}) \doteq 0$ for $(k, l) \in \{(1, 0), (1, 1), (0, 1)\}$. For the terms t_{10} and t_{01}

(i.e., the marginal first moments), it can thus be achieved with marginal random hot-deck imputation. However, the cross-product term, t_{11}, is more problematic since it is a measure of the relationship between x and y. Marginal imputation, which consists of imputing x and y separately, tends to attenuate the relationship between variables and, as a result, introduces a bias that may be severe if the nonresponse rate is appreciable. To deal with this issue, survey statisticians typically use an alternative version of random hot-deck imputation, which consists of selecting a common donor at random from the set s_{rr} of common donors (i.e., the set of sample units that responded to both items) when both items are missing. Unfortunately, although this imputation procedure generates less bias than marginal random hot-deck imputation, it does not succeed in eliminating it completely, unless $s_{rm} = s_{mr} = \emptyset$. This point is further discussed in Section 3.1.

3 Imputation procedures

In this section, we describe three random imputation procedures: (i) random hot-deck imputation (ii) joint random hot-deck imputation and (iii) balanced joint random hot-deck imputation. For each procedure, the asymptotic bias of $\hat{\rho}_{xyI}$ is examined.

3.1 Random hot-deck imputation procedure

Random hot-deck imputation may be described as follows:

(i) for $i \in s_{mr}$, missing x_i is imputed by $x_i^* = x_j, j \in s_{rr} \cup s_{rm}$ such that

$$P\left(x_i^* = x_j\right) = \frac{w_j}{\sum_{k \in s} w_k r_{xk}};$$

(ii) for $i \in s_{rm}$, missing y_i is imputed by $y_i^* = y_j, j \in s_{rr} \cup s_{mr}$ such that

$$P\left(y_i^* = y_j\right) = \frac{w_j}{\sum_{k \in s} w_k r_{yk}};$$

(iii) for $i \in s_{mm}$, missing (x_i, y_i) is imputed by $\left(x_i^*, y_i^*\right) = \left(x_j, y_j\right), j \in s_{rr}$ such that

$$P\left\{\left(x_i^*, y_i^*\right) = \left(x_j, y_j\right)\right\} = \frac{w_j}{\sum_{k \in s} w_k r_{xk} r_{yk}}.$$

Under this imputation procedure, the relative conditional nonresponse bias of $\hat{\rho}_{xyI}$, $RB_{qI}\left(\hat{\rho}_{xyI}\right) = B_{qI}\left(\hat{\rho}_{xyI}\right)/\hat{\rho}_{xy\pi}$, can be approximated by

$$RB_{qI}\left(\hat{\rho}_{xyI}\right) \doteq -\left(1 - p_{rr} - p_{mm}\right), \tag{4}$$

provided $\hat{\rho}_{xy\pi} \neq 0$; see Chauvet and Haziza (2011). If $\hat{\rho}_{xy\pi} = 0$ (i.e, the variables x and y are unrelated), the imputed estimator $\hat{\rho}_{xyI}$ is asymptotically qI-unbiased for $\hat{\rho}_{xy\pi}$, as expected. Expression (4) shows that the asymptotic bias is always negative and that it vanishes if $p_{rm} = p_{mr} = 0$, or equivalently, if $s_{rm} = s_{mr} = \emptyset$. In general however, this imputation distorts the relationship between x and y.

3.2 Joint imputation procedure

In this section, we introduce a simple joint imputation procedure. Let $1(.)$ be the usual indicator function. The proposed method may be described as follows:

(i) for $i \in s_{mr}$, $\epsilon \in \{0,1\}$, missing x_i is imputed by $x_i^* = x_j, j \in s_{rr}$ such that

$$P\left(x_i^* = x_j \left| y_i = \epsilon\right.\right) = \begin{cases} \frac{w_j}{\sum_{k\in s} w_k r_{xk} r_{yk} 1(y_k=\epsilon)} & \text{if } y_j = \epsilon \\ 0 & \text{otherwise;} \end{cases}$$

(ii) for $i \in s_{rm}$, $\epsilon \in \{0,1\}$, missing y_i is imputed by $y_i^* = y_j, j \in s_{rr}$ such that

$$P\left(y_i^* = y_j \left| x_i = \epsilon\right.\right) = \begin{cases} \frac{w_j}{\sum_{k\in s} w_k r_{xk} r_{yk} 1(x_k=\epsilon)} & \text{if } x_j = \epsilon \\ 0 & \text{otherwise;} \end{cases}$$

(iii) for $i \in s_{mm}$, missing (x_i, y_i) is imputed by $(x_i^*, y_i^*) = (x_j, y_j), j \in s_{rr}$ such that

$$P\left\{(x_i^*, y_i^*) = (x_j, y_j)\right\} = \frac{w_j}{\sum_{k\in s} w_k r_{xk} r_{yk}}.$$

It is shown in the Appendix that $B_{qI}(\hat{\rho}_{xyI}) \doteq 0$ under this imputation procedure. A drawback of the proposed procedure is that it suffers from an additional variability, called the imputation variance, due to the random selection of the donors. As a result, it is not fully efficient, a term coined by Kim and Fuller (2004).

3.3 Balanced joint imputation procedure

To eliminate the imputation variance, we suggest selecting the donors at random so that the imputation error in (3), $\hat{\rho}_{xyI} - \tilde{\rho}_{xyI}$, is equal to zero. It suffices to select the donors so that the following constraints are satisfied:

$$\hat{t}_{kl,I} = E_I(\hat{t}_{kl,I} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}), \tag{5}$$

for any $(k,l) \in \{(1,0),(1,1),(0,1)\}$. An imputation procedure that satisfies (5) has been called *balanced*, a term coined by Chauvet, Deville and Haziza (2011).

We first introduce some further notation. Let

$$\hat{p}_{kl} = \frac{\sum_{m\in s} w_m r_{xm} r_{ym} 1(x_m = k) 1(y_m = l)}{\sum_{m\in s} w_m r_{xm} r_{ym}}$$

for any $(k,l) \in \{(0,0),(1,0),(0,1),(1,1)\}$, which represents the probability of imputing $(x_i^*, y_i^*) = (k,l)$ when $i \in s_{mm}$. Also, for $\epsilon \in \{0,1\}$ let

$$\hat{p}_{x|y=\epsilon} = \frac{\sum_{m\in s} w_m r_{xm} r_{ym} x_m 1(y_m = \epsilon)}{\sum_{m\in s} w_m r_{xm} r_{ym} 1(y_m = \epsilon)}$$

which represents the probability of imputing $x_i^* = 1$ when $i \in s_{mr}$ and $y_i = \epsilon$, and

$$\hat{p}_{y|x=\epsilon} = \frac{\sum_{m \in s} w_m r_{xm} r_{ym} y_m 1(x_m = \epsilon)}{\sum_{m \in s} w_m r_{xm} r_{ym} 1(x_m = \epsilon)}$$

which represents the probability of imputing $y_i^* = 1$ when $i \in s_{rm}$ and $x_i = \epsilon$.

Let us first consider the case $(k, l) = (1, 0)$. We have

$$\begin{aligned} \hat{t}_{10,I} &= \sum_{i \in s} w_i r_{xi} x_i + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i x_i^* \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} (1 - y_i) x_i^* + \sum_{i \in s} w_i (1 - r_{xi})(1 - r_{yi}) x_i^*. \end{aligned} \tag{6}$$

Furthermore, we have

$$\begin{aligned} E_I \left\{ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i x_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r} \right\} &= \hat{p}_{x|y=1} \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i, \\ E_I \left\{ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} (1 - y_i) x_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r} \right\} &= \hat{p}_{x|y=0} \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} (1 - y_i), \\ E_I \left\{ \sum_{i \in s} w_i (1 - r_{xi})(1 - r_{yi}) x_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r} \right\} &= (\hat{p}_{10} + \hat{p}_{11}) \sum_{i \in s} w_i (1 - r_{xi})(1 - r_{yi}). \end{aligned} \tag{7}$$

From (6) and (7), it follows that $\hat{t}_{10,I} = E_I(\hat{t}_{10,I} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})$ if the following balancing equations are satisfied:

$$\begin{aligned} \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i (x_i^* - \hat{p}_{x|y=1}) &= 0, \\ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} (1 - y_i)(x_i^* - \hat{p}_{x|y=0}) &= 0, \\ \sum_{i \in s} w_i (1 - r_{xi})(1 - r_{yi})(x_i^* - \hat{p}_{10} - \hat{p}_{11}) &= 0. \end{aligned}$$

Similar balancing equations may be derived for the cases $(k, l) = (0, 1)$ and $(k, l) = (1, 1)$ in (5). After some algebra, we obtain that the constraints in (5) are satisfied if the imputation procedure is such that (i) the two balancing equations (corresponding to imputation on s_{mr})

$$\sum_{i \in s_{mr}} w_i \left\{ y_i (x_i^* - \hat{p}_{x|y=1}), (1 - y_i)(x_i^* - \hat{p}_{x|y=0}) \right\} = 0$$

are satisfied; (ii) the two balancing equations (corresponding to imputation on s_{rm})

$$\sum_{i \in s_{rm}} w_i \left\{ x_i (y_i^* - \hat{p}_{y|x=1}), (1 - x_i)(y_i^* - \hat{p}_{y|x=0}) \right\} = 0$$

are satisfied and (iii) the three balancing equations (corresponding to imputation on s_{mm})

$$\sum_{i \in s_{mm}} w_i \left\{ x_i^* - \hat{p}_{10} - \hat{p}_{11}, y_i^* - \hat{p}_{01} - \hat{p}_{11}, x_i^* y_i^* - \hat{p}_{11} \right\} = 0$$

are satisfied. In other words, we perform the imputations separately on each of the subsamples s_{mr}, s_{rm} and s_{mm}.

4 Simulation study

We conducted a limited simulation study to test the performance of the procedures described in Section 3. We first generated 3 finite populations of size $N = 10,000$, each containing two binary variables of interest x and y. The variables x and y were generated to obtain a population coefficient of correlation ρ equal to 0.3 for population 1, 0.5 for population 2 and 0.7 for population 3. We were interested in estimating the marginal first moments p_{10} and p_{01} as well as the population correlation coefficient ρ_{xy} given by (1).

From each population, we selected $B = 1000$ samples of size $n = 500$ according to simple random sampling without replacement. Then, in each selected sample, nonresponse to items x and y was generated according to

$$P\left(r_{xi} = \epsilon, r_{yi} = \eta\right) = \left(p_{rr}\right)^{\epsilon\eta} \left(p_{rm}\right)^{\epsilon(1-\eta)} \left(p_{mr}\right)^{(1-\epsilon)\eta} \left(p_{mm}\right)^{(1-\epsilon)(1-\eta)} \qquad (8)$$

with $\epsilon \in \{0,1\}$ and $\eta \in \{0,1\}$. We used three configurations, which we call mechanisms 1, 2 and 3, of the vector $\left(p_{rr}, p_{rm}, p_{mr}, p_{mm}\right)'$. We used

$$\left(p_{rr}, p_{rm}, p_{mr}, p_{mm}\right) = \begin{cases} (0.2, 0.25, 0.25, 0.3) & \text{for mechanism 1,} \\ (0.4, 0.15, 0.15, 0.3) & \text{for mechanism 2,} \\ (0.6, 0.05, 0.05, 0.3) & \text{for mechanism 3.} \end{cases}$$

We computed the imputed estimators of the marginal first moments $\hat{p}_{10,I}$ and $\hat{p}_{01,I}$ and the imputed estimator of the correlation coefficient, $\hat{\rho}_{xyI}$, based on (i) the random hot-deck imputation (RHDI) procedure described in Section 3.1; (ii) the proposed joint random hot-deck imputation (JHDI) procedure described in Section 3.2 and (iii) the proposed balanced random hot-deck imputation (BHDI) procedure described in Section 3.3. Also, we computed the estimators of the marginal first moments and of the correlation coefficient based on the complete cases (CC). Note that unlike the imputed estimators, the CC estimators require the response flags to be available in the imputed data file and can not be computed if these flags are not available.

As a measure of bias of a point estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo Percent Relative Bias (RB) given by

$$RB(\hat{\theta}) = \frac{E_{MC}(\hat{\theta}) - \theta}{\theta} \times 100, \qquad (9)$$

where $E_{MC}(\hat{\theta}) = B^{-1} \sum_{b=1}^{B} \hat{\theta}^{(b)}$ and $\hat{\theta}^{(b)}$ denotes the estimator $\hat{\theta}$ in the b-th sample, $b = 1, \ldots, 1000$.

Table 1 shows the Monte Carlo percent Relative Bias (RB) corresponding to $\hat{p}_{10,I}$, $\hat{p}_{01,I}$ and $\hat{\rho}_{xyI}$. For the marginal first moments p_{10} and p_{01}, all the imputation procedures and the complete case estimator led to negligible bias, as expected. Turning to the coefficient of correlation ρ_{xy}, we note that its imputed estimator $\hat{\rho}_{xyI}$, was considerably biased under RHDI in all the scenarios. Also, the bias was negative clearly illustrating the problem of attenuation of relationships under this type of imputation procedure. This is consistent with the bias expression (4). On the other hand, both JHDI and BHDI led to negligible bias, showing that both procedures succeeded in preserving the relationship between variables. Also, CC led to negligible bias as expected.

Table 1: Monte-Carlo Relative Bias of the estimators

	Population 1			Population 2			Population 3		
	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$
				Mechanism 1					
CC	0.00	-0.25	0.13	0.03	-0.07	0.76	-0.19	-0.23	-0.14
RHDI	0.06	-0.19	-49.64	-0.10	-0.16	-49.61	-0.25	-0.21	-50.20
JHDI	-0.10	-0.10	0.28	0.02	-0.12	0.88	-0.19	-0.20	0.00
BHDI	-0.09	-0.08	0.39	0.03	-0.10	0.98	-0.24	-0.19	0.04
				Mechanism 2					
CC	-0.11	-0.04	-0.24	0.22	0.13	-0.13	-0.13	0.09	0.21
RHDI	-0.05	0.00	-30.37	0.30	0.12	-30.17	-0.03	0.14	-29.93
JHDI	-0.03	0.06	-0.89	0.29	-0.01	-0.40	-0.07	0.07	0.42
BHDI	-0.06	-0.01	-0.16	0.29	0.08	-0.04	-0.04	0.01	0.25
				Mechanism 3					
CC	-0.29	0.11	0.41	0.00	-0.17	-0.46	0.15	0.01	-0.10
RHDI	-0.27	0.17	-10.10	-0.01	-0.15	-10.64	0.11	0.04	-10.15
JHDI	-0.28	0.12	0.38	0.03	-0.16	-0.71	0.19	0.02	-0.07
BHDI	-0.29	0.10	0.45	0.01	-0.21	-0.46	0.18	-0.03	-0.08

We then compared the efficiency of JHDI and BHDI. Let $\hat{\theta}^{JHDI}$ and $\hat{\theta}^{BHDI}$ denote the estimator $\hat{\theta}$ under JHDI and BHDI, respectively. As a measure of Relative Efficiency (RE), we used

$$RE = \frac{MSE_{MC}(\hat{\theta}^{(\cdot)})}{MSE_{MC}(\hat{\theta}^{(JHDI)})}, \tag{10}$$

$MSE_{MC}(\hat{\theta})$ is the Monte Carlo mean square error of $\hat{\theta}$.

Table 2 shows the RE corresponding to $\hat{t}_{10,I}$, $\hat{t}_{01,I}$ and $\hat{\rho}_{xyI}$. It is clear that the imputed estimators under BHDI were significantly more efficient than those obtained under JHDI in all the scenarios, with values of RE ranging from 0.78 to 0.87. The CC estimators were also more efficient than the imputed estimators under JHDI in all the scenarios, with values of RE ranging from 0.64 to 0.89. For mechanisms 2 and 3, BHDI and CC led to very similar results. For mechanism 1, the CC estimators were more efficient than those obtained under BHDI. These results can be explained by the fact that, since $p_{rr} = 0.2$ for mechanism 1, the number of common donors was rather small. Consequently, both JHDI and BHDI used only a subset of the set of common donors in order to perform the imputations, which contributed to the variability of the imputed estimators. On the other hand, the CC estimators were computed using all the units in the set of common donors.

5 Application to the French Wealth Survey

In this section, we apply the proposed imputation procedures in the context of the FWS.

Table 2: Monte-Carlo Relative Efficiency of the imputed estimators

	Population 1			Population 2			Population 3		
	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$	$\hat{p}_{10,I}$	$\hat{p}_{01,I}$	$\hat{\rho}_{xyI}$
				Mechanism 1					
CC	0.67	0.66	0.86	0.67	0.64	0.84	0.73	0.80	0.86
BHDI	0.86	0.87	0.86	0.85	0.80	0.85	0.83	0.87	0.86
				Mechanism 2					
CC	0.78	0.77	0.81	0.78	0.81	0.81	0.84	0.88	0.78
BHDI	0.82	0.83	0.81	0.80	0.82	0.82	0.80	0.83	0.78
				Mechanism 3					
CC	0.78	0.81	0.78	0.83	0.81	0.81	0.82	0.89	0.79
BHDI	0.79	0.82	0.78	0.84	0.80	0.82	0.80	0.87	0.80

5.1 Methodology of the French Wealth Survey

To reduce nonresponse as much as possible, the FWS questionnaire is divided into two main parts. The first part draws the inventory of all the assets held by the interviewed household (e.g., financial assets, housing wealth, business wealth, indebtedness). Item nonresponse was rather uncommon for the variables included in the first part, concerning twenty or so households. The second part collects a detailed description of the listed assets (e.g., securities accounts' characteristics, life annuities' characteristics, number of assets of each kind, amount on each asset). The variables included in the second part were heavily prone to item nonresponse, due to greater difficulties for the households to describe the assets.

It is well known that the distribution of most variables related to wealth are highly skewed since only a small fraction of the French households are very wealthy. In order to obtain accurate estimates, it is thus important to interview enough "wealthy households". To that end, the population was stratified using some auxiliary variables related to the variables of interest. For metropolitan France and Reunion, tax income registers were used as the sampling frame. For the French West Indies, the samples of the New Census were used as the sampling frame since the corresponding tax income registers were judged to be insufficiently reliable.

At the end of the data collection stage, approximately 20,000 households were selected in the sample and approximately 15,000 households completed the questionnaire to its very end. The FWS weighting process can be described as follows: to compensate for unit non-response, the basic weights (which are defined as the inverse of the sample inclusion probabilities) of the responding households were adjusted using the inverse of the estimated response probabilities. These estimated probabilities were computed within weighting classes formed on the basis of auxiliary variables available for both responding and nonresponding households. The resulting adjusted weights were further adjusted using a calibration procedure. The calibration variables included the type of location of the dwelling, the age, degree and occupation of the household's reference person, the type of family, the income of the household and the number of people of each age bracket and gender.

Imputation was performed independently within classes. Three imputation classes were used: the first class consisted of overseas population, the second one of wealthy individuals, whereas the third one consisted of the remaining individuals.

5.2 Imputation of the securities accounts' characteristics

In France, two types of securities accounts are distinguished: (i) the comptes-titres (CT), that is, classical accounts which enable owning bonds, stocks and mutual funds; (ii) the Plans Epargnes en Actions (PEA), which are special accounts limited to European Union investments and profit by a fiscal allowance. Note that an individual cannot own more than one PEA.

One of the important objectives of the FWS lies in estimating the proportion of households owning each type of assets, and in particular, the proportion of households owning stocks. The presence (and possibly, the proportion) of stocks is also a measure of the true risk involved by the detention of the securities account. In addition, the households are also asked about the risk tolerance of their securities accounts, which measures, to a certain extent, the perception of the risk they incur by holding this asset. The incurred risk is important to determine whether the recent financial crisis has curbed down the propensity to invest in risky assets. Moreover, the link between the part of stocks and the risk tolerance measures the gap between the perception and the reality of the incurred risk.

The variable of stock owning consists in two categories: with stocks, or without stocks. The risk tolerance consists in three categories: no risk, medium risk and high risk. Due to a bad implementation of a filter question, none of the households was asked the risk tolerance for their CT, whereas it was asked for the PEAs. For imputation purposes, both types of securities accounts (CT and PEA) were put together, in order to use the characteristics of the PEA to impute the CT.

The variables of stock owning and the risk tolerance of both types of securities accounts were imputed jointly. Due to a high nonresponse rate (among the $7,073$ securities accounts in the sample, there were $3,562$ missing values for risk tolerance only, 113 for presence of stocks only, and 77 for both), we focused on whether or not the securities account are risky. Thus, the risk tolerance was recoded into two categories, indicating if the securities account is risky or not.

To fill in the missing values, the three imputation methods described in Section 3 were used. In addition, we computed the complete case estimates which are based on the responding units only. The results are shown in Table 3. We note that the complete case estimates as well as JHDI and BHDI led to similar results with correlation values ranging between 15.9% and 17.9%. On the other hand, the coefficient of correlation under RHDI was substantially smaller with a value equal to 7.8%, illustrating the attenuation of relationships.

Table 3: Proportion of securities accounts with stocks, proportion of risky accounts and coefficient of correlation between stock owning and risk, with estimation based on complete cases or on three imputation procedures

	Percent of accounts		Coefficient of correlation
	with stocks	with risk	
Complete case	0.749	0.210	0.167
RHDI	0.729	0.203	0.078
JHDI	0.731	0.202	0.159
BHDI	0.728	0.203	0.179

5.3 Imputation of life-annuities' characteristics

Life-annuities is the most common type of risky assets in France, with approximately 34.5 % of French households owning one. In what follows, we focus on life-annuities that contain market assets.

The variable giving the proportion of stocks in the investment consists in five categories: less than one third (A), between one third and one half (B), between one half and two thirds (C), more than two thirds and less than one (D), equal to one (E). The variables giving the part of stocks in the investment and the risk tolerance (in three categories) are imputed jointly. Among the 4,832 life annuities containing market assets, there were 784 missing values for the risk tolerance, 93 for the part of stocks and 510 for both. These rates of missing information were considered low enough to keep the risk tolerance into three categories.

To fill in the missing values, we used RHDI, JHDI and BHDI, see Section 3. Table 4 shows the estimated proportions and Spearman's rank coefficient of correlation. As before, we note that the complete case estimate, JHDI and BHDI led to very similar results in all the scenarios. On the other hand, the estimate obtained under RHDI imputation was significantly smaller.

6 Concluding remarks

In this paper, we considered the problem of preserving the relationship between categorical variables when imputation was used to compensate for the missing values. We proposed a simple joint imputation procedure that succeeds in preserving the relationship between two categorical variables, unlike random hot-deck imputation. We also proposed a fully efficient version of the proposed joint imputation procedure. Simulation results clearly demonstrated the good performance of both methods in terms of bias. Also, the balanced random hot-deck imputation procedure was found to be significantly more efficient than the joint random hot-deck imputation procedure.

Table 4: Proportion of life-annuities classified by proportion of stocks, proportion of life-annuities classified by risk and Spearman's rank coefficient of correlation, with estimation based on the common respondents and on three imputation procedures

	Percent of life-annuities with a proportion of stocks in the investment					Percent of life annuities with			Coefficient of correlation
	A	B	C	D	E	no risk	medium risk	high risk	
Complete case	52.2	19.9	12.0	6.0	10.0	58.7	26.3	15.0	0.518
RHDI	51.4	20.5	11.9	6.2	10.0	60.0	26.4	13.6	0.434
JHDI	52.1	20.5	12.0	5.7	9.8	60.0	26.1	13.9	0.515
BHDI	53.2	19.5	11.7	6.1	9.6	60.6	25.6	13.8	0.508

References

Chauvet, G., Deville, J.C., and Haziza, D. (2011). On Balanced Random Imputation in Surveys. *Biometrika*, 98, 459-471.

Chauvet, G., and Haziza, D. (2011). Fully efficient estimation of coefficients of correlation in the presence of imputed data. *To appear in the Canadian Journal of Statistics.*

Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, Editors: C.R. Rao and D. Pfeffermann, 215-246.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, 91, 559-578.

Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association*, 97, 544-552.

Skinner, C. J. and Rao, J. N. K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*, 102, 149-167.

Appendix: Asymptotic unbiasedness under the joint imputation procedure

In this section, we show that $B_{qI}\left(\hat{\rho}_{xyI}\right) \doteq 0$. To that end, we need to show that $B_{qI}\left(\hat{t}_{kl,I}\right) \doteq 0$ for $(k,l) \in \{(1,0),(1,1),(0,1)\}$. We start by showing that $E_{qI}\left(\hat{t}_{10,I}\right) \doteq \hat{t}_{10,\pi}$. The proof corresponding to $\hat{t}_{01,I}$ is similar. It follows from (6) and (7) that

$$\begin{aligned} \tilde{t}_{10,I} &= \sum_{i\in s} w_i r_{xi} x_i + \hat{p}_{x|y=1} \sum_{i\in s} w_i (1-r_{xi}) r_{yi} y_i \\ &+ \hat{p}_{x|y=0} \sum_{i\in s} w_i (1-r_{xi}) r_{yi} (1-y_i) + (\hat{p}_{10}+\hat{p}_{11}) \sum_{i\in s} w_i (1-r_{xi})(1-r_{yi}). \end{aligned}$$

Taking expectation with respect to the nonresponse model, we obtain

$$\begin{aligned} E_q\left(\tilde{t}_{10,I}|\mathbf{x},\mathbf{y},\boldsymbol{\delta}\right) &\doteq (p_{rr}+p_{rm}) \sum_{i\in s} w_i x_i \\ &+ p_{mr} \sum_{i\in s} w_i x_i y_i + p_{mr} \sum_{i\in s} w_i x_i (1-y_i) \\ &+ p_{mm} \left\{ \sum_{i\in s} w_i x_i (1-y_i) + \sum_{i\in s} w_i x_i y_i \right\} \\ &= (p_{rr}+p_{rm}+p_{mr}+p_{mm}) \sum_{i\in s} w_i x_i \\ &= \hat{t}_{10,\pi}. \end{aligned}$$

We now show that $E_{qI}\left(\hat{t}_{11,I}\right) \doteq \hat{t}_{11,\pi}$. First, we write $\hat{t}_{11,I}$ as

$$\begin{aligned} \hat{t}_{11,I} &= \sum_{i\in s} w_i r_{xi} r_{yi} x_i y_i \\ &+ \sum_{i\in s} w_i (1-r_{xi}) r_{yi} y_i x_i^* + \sum_{i\in s} w_i (1-r_{yi}) r_{xi} x_i y_i^* \\ &+ \sum_{i\in s} w_i (1-r_{yi})(1-r_{xi}) x_i^* y_i^*. \end{aligned}$$

It follows from (7) and

$$\begin{aligned} E_I\left\{ \sum_{i\in s} w_i (1-r_{yi}) r_{xi} x_i y_i^* |\mathbf{x},\mathbf{y},\boldsymbol{\delta},\mathbf{r} \right\} &= \hat{p}_{y|x=1} \sum_{i\in s} w_i (1-r_{yi}) r_{xi} x_i, \\ E_I\left\{ \sum_{i\in s} w_i (1-r_{xi})(1-r_{yi}) x_i^* y_i^* |\mathbf{x},\mathbf{y},\boldsymbol{\delta},\mathbf{r} \right\} &= \hat{p}_{11} \sum_{i\in s} w_i (1-r_{xi})(1-r_{yi}) \end{aligned}$$

that

$$\begin{aligned} \tilde{t}_{11,I} &= \sum_{i\in s} w_i r_{xi} r_{yi} x_i y_i \\ &+ \hat{p}_{x|y=1} \sum_{i\in s} w_i (1-r_{xi}) r_{yi} y_i + \hat{p}_{y|x=1} \sum_{i\in s} w_i (1-r_{yi}) r_{xi} x_i \\ &+ \hat{p}_{11} \sum_{i\in s} w_i (1-r_{xi})(1-r_{yi}). \end{aligned}$$

This leads to

$$
\begin{aligned}
E_q\left(\tilde{t}_{11,I}|\mathbf{x},\mathbf{y},\boldsymbol{\delta}\right) &\doteq p_{rr}\sum_{i\in s} w_i x_i y_i \\
&+ p_{mr}\sum_{i\in s} w_i x_i y_i + p_{rm}\sum_{i\in s} w_i x_i y_i \\
&+ p_{mm}\sum_{i\in s} w_i x_i y_i \\
&= \hat{t}_{11,\pi}.
\end{aligned}
$$